UNIVERSITI TEKNOLOGI MARA

University Publication Centre (UPENA)

Journal of Mechanical Engineering

An International Journal

Volume 6 No. 1 **April 2009** **ISSN 1823-5514**

JOURNAL OF MECHANICAL ENGINEERING
(JMechE)

Journal of Mechanical Engineering is jointly published by the Faculty of Mechanical Engineering and University Publication Centre (UPENA), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.

Journal of Mechanical Engineering

An International Journal

Volume 6 No. 1 **April 2009** **ISSN 1823-5514**

A Method for Determining Tool Group Flexibility with Uncertain Machine Availability – Applications in a Semiconductor Manufacturing Process

Adam Terry[1], M. Ramulu[2], & P.N. Rao[3]
[1] Intel Corporation, Portland, OR

[2] Program in Engineering and Manufacturing Management, University of Washington

[3] Department of Industrial Technology, University of Northern Iowa, Cedar Falls, IA

ABSTRACT

The production of Integrated Circuits (IC) is a detailed and exacting process requiring tight specifications and precise equipment. The high cost and unique traits of this equipment requires high utilization and maximum throughput to achieve real profits. The design of fabrication facility (FAB) processes requires a thorough understanding of the adverse effects that random machine availability has on system performance. These effects (increased cycle time, decreased and variable throughput, etc) can be offset by tool group flexibility. Tool group flexibility can be described by two measures: machine flexibility (the number of tasks a machine can perform) and task flexibility (the number of machines qualified to perform a specific task). These two measures are related by the ratio of the number of machines in the tool group to the number of tasks that the group must perform. This paper utilizes a combined linear programming and simulation approach in an attempt to model the manufacturing system to gain insight into the production dynamics. The model is based on current production methodology and the use of modular equipment (steppers). The results include some insight into the added cost of flexibility and the associated production ramifications.

Keywords: *Linear Programming (LP), Simulation, Integrated Circuit, Fabrication, Re-entrant systems, modeling, semiconductor, machine flexibility*

Introduction

IC manufacturing processes can be divided into four main steps: wafer fabrication, wafer probe, assembly, and final testing. Wafer manufacturing primarily deals with taking polished silicon wafers (up to 300 mm in diameter) and fabricating a grid of hundreds of individual dies (IC) on the wafer surface. Each wafer is cut into the individual die and tested electrically by means of thin probes. Assembly involves placing the IC into protective plastic or ceramic packaging. The product is then sent to final testing to ensure the customers receive a product that meets specifications.

Fabrication is by far the most challenging step in IC manufacturing, and more advanced IC products (such as microprocessors) often require a complicated combination of hundreds of individual processes to produce a finished fabricated wafer. However, all of these wafer fabrication steps can be summarized by five main operations, namely, Layering, Patterning, Planarization, Doping, and Heat Treating [1]. These five main operations create patterned layers on a silicon wafer. Initial layers are patterned to form the transistor on the wafer surface, and then subsequent layers are patterned to form the circuitry that connects the transistors. The completed product often requires more than 20 patterned layers. Currently, one of the main technological constraints in semiconductor fabrication is lithography. Although the principle behind lithography is simple, it is difficult to effectively perform on a sub-micron scale. Consequently, very expensive stepper machines are required to do the lithography tasks.

Given the challenges of semiconductor fabrication, a significant amount of effort is required to develop and continuously improve fabrication technologies and facilities. A semiconductor manufacturer will invest large amounts of capital and employ a sizeable workforce to profitably fabricate wafers. Each new FAB can cost in excess of $2 Billion, with over 80% of the total FAB construction costs attributed to equipment [3]. As the investment required to fabricate wafers increases (and the ratio of fixed costs to variable costs of producing wafers increases), the impact of properly managing FAB capacity becomes increasingly important and contributes directly to the financial bottom line.

While this environment creates significant pressure for FABs to operate at maximum capacity, the dangers of overloading a system are well known. In the short life cycle of processors, long cycle times and missed due dates from improper management of FAB capacity can result in significant losses. In addition to longer-range capacity planning, shorter-term scheduling on a daily or weekly basis adds an additional level of complexity in keeping the FAB output optimal. Operational decisions regarding the quantity and timing of new wafers into the FAB (wafer release policies) and how to control the resources and wafers currently in the FAB (resource scheduling) are extremely complex.

There are a number of characteristics in semiconductor manufacturing that make capacity planning and scheduling difficult [4]. In particular:

- Complex and Re-entrant Product Flows: Wafers will return to the same group of machines several times throughout production, resulting in re-entrant process flow. Wafers that are at different production stages compete for service from the same group of machines.
- Unreliable Equipment: Most critical equipment is extremely sensitive and subject to random failures and unpredictable performance, despite extensive calibration and maintenance routines.

Quality issues can have a dramatic affect on capacity planning. The quality of finished wafers is largely dependent on the die yield (the number of dies on the wafer that work), and the performance of those dies. Variations in production quality cause some dies to perform better than others, while some dies are entirely dysfunctional.

Frequent testing is essential due to the extremely precise processing requirements and exceptionally long production time (it will often take several weeks for a wafer to complete fabrication). Line yield is a measure of the number of wafers that complete processing vs. the number that started. Rework, line yield, and die yield can significantly alter the capacity requirements within the FAB. Therefore, the real throughput of the FAB is determined by the number of dies that meet performance requirements in addition to the number of wafers that are finished:

$$\text{Throughput} = (\text{Wafer Starts})\,(\text{Line Yield})\,(\text{Die Yield})$$

An important process variation issue in the lithography tool group is eliminated by stepper-wafer dedication. Stepper-wafer dedication requires a wafer to be processed by the same stepper each time it returns for a lithography patterning operation. The precision required by the IC design exposes a consistent but specific pattern distortion "signature" that each stepper leaves on the layer. In order to match the next layer correctly over the previous, the layer must be patterned by the same stepper so that consistent pattern distortion occurs [5].

Steppers have other relevant issues in capacity planning. First, stepper availability is extremely important. The complexity of the steppers make them susceptible to failure, necessitating an extensive maintenance and calibration schedule. The frequency and duration of failures and duration of scheduled downtimes are unpredictable and can significantly reduce potential capacity.

Inefficiencies in the stepper tool group translate into decreased FAB performance. Understanding the adverse effects that random machine availability has on system performance is essential. The main effects that will be considered in this paper are:

- Decreased Throughput – Stepper tool group capacity is underutilized every time a stepper goes idle while WIP piles up in front of a failed stepper.

Workload imbalances within the stepper tool group can cause a single layer to become the constraint in the tool group. Throughput is maximized in a balanced tool group.

- Increased Cycle Time – Whenever a wafer lot waits to be processed, the cycle time and the amount of work in process (WIP) in the system increases.
- Increased Variability in Throughput – Although a FAB may perform to goals on a long-term average, variability in throughput (and cycle time) on a week-to-week basis is detrimental. FABs often operate on a weekly shipment schedule, and meeting customer expectations is paramount.

Flexibility is an ambiguous term that can have a number of different definitions and interpretations [6]. This study defines tool group flexibility with two measures: Machine (stepper) flexibility and task (layer) flexibility. *Machine flexibility* is defined as the number of tasks each machine can perform. In lithography terms, stepper flexibility is the number of layers that a stepper is qualified to process (layers per stepper, or LS). As the LS for each stepper is increased, the following associations can be assumed:

- Higher Quality: A high LS ratio does not guarantee higher quality, and stepper-wafer dedication is impossible without being qualified for all of the critical layers.
- Higher Cost/Complexity: Qualifying a machine to process additional layers requires routine maintenance, increased complexity in scheduling and coordination, and requires additional reticles to be bought and managed.
- Better Line Balance and Higher Stepper Utilization: A higher LS increases the ability of a stepper to switch from one layer to another depending on WIP levels.

Task flexibility is defined as the number of machines that can perform each individual task. In lithography terms, layer flexibility is the number of steppers that are qualified to process each layer (steppers per layer, or SL). As the SL for each layer is increased there will be higher potential capacity for the layer at any given time. A higher SL means that more machines will be qualified to process the layer.

The relationship between stepper and layer flexibility depends on the ratio of number of steppers (N) to number of layers (M) such that LS = SL/(N/M). If the number of steppers and tasks within the tool group are fixed, then the tool group flexibility is defined as the number of layers each tool is assigned (qualified) to process.

For example, for a tool group with 5 steppers and 10 layers, the N/M ratio = 0.5. In the minimal flexibility case, each layer would be qualified to be processed on 1 stepper, sharing stepper capacity with another layer. Therefore, SL = 1 and LS = 2. If a stepper failed, the queues for both layers would continue to build until the stepper became available again. However, by

increasing layer flexibility to SL = 2 (qualifying another stepper to be able to process these layers) the queue buildup could be averted. When one stepper fails for a prolonged period, the second stepper could switch from processing the downstream layers to the upstream layer instead of going idle while waiting for upstream wafers to be processed.

This paper focuses on developing an approach to answering a basic capacity planning issue for the lithography tool group: given a certain number of steppers and layers (machines and tasks), which steppers should be qualified to process which layers?

A fundamental aspect is to understand the effects that stepper availability and tool group flexibility have on the key FAB performance metrics. Therefore, the purpose of this paper is to develop a methodology for determining an optimal level of tool group flexibility. Ultimately, the methodology should quantitatively establish the relationship between FAB performance and different levels of tool group flexibility. These results can then be used to make educated flexibility/dedication decisions as additional costs are weighed against increasing corresponding FAB throughput and cycle time performance.

To achieve this objective, the four main goals of this study are:

- Create a deterministic linear programming (LP) model that provides stepper-layer qualification assignments to maximize throughput.
- Create a simulation that tests the LP solution under a stochastic environment.
- Develop an algorithm that integrates the LP and Simulation tools, providing a standard method of evaluating the relationship between FAB performance and different levels of tool group flexibility.
- Test and verify the methodology with a simplified representation of a lithography tool group.

Literature Review

Fowler et al. [7] uses a simulation-based analysis to examine cycle-time in an existing semiconductor FAB with the goal of finding potential areas for improvement. The major finding of their research, though not explicitly explored, suggests that increased tool flexibility could significantly improve FAB performance. Hase et al [8] provide an excellent analysis of five different machine layout/dedication policies and their effect on cycle time. Peikert et al. [10] demonstrate the method of simulating just the lithography tool group in a semiconductor FAB to rapidly model the effects of lithography operational decisions. Kumar [2] discusses re-entrant lines in queuing network terms and emphasizes the importance of considering machine failures and setup times.

Benjaafar and Gupta [14] explore workload allocation to minimize work in process (WIP) while considering setup costs. Leachman and Carmon [15] study

resource sharing between different operations, emphasizing task performance differences between machine types. Rohan [16] develops a workload allocation algorithm that allocates resources to tasks in order to predict utilizations. Rohan [17] explores the two extremes of entirely flexible tool groups or entirely dedicated tool groups and the decision criteria on choosing between the two. He proposes that dedication decisions can be based on two ratios: the magnitude of setup time relative to process time and the N/M (number of machines/number of tasks) ratio.

This research is unique in that it develops a methodology of determining an appropriate level of tool group flexibility on a machine-task level. The methodology addresses the characteristics of re-entrant flow and the lithography stepper group. Previous work has addressed various aspects of flexibility on this level of detail, but does not incorporate re-entrant flow and lithography tool group considerations.

Model Formulation

In the scenario that prompted this study, the high-volume capacity of the lithography tool group for a new FAB is being planned. The total number of steppers that will be used in the group has already been determined by preliminary capacity planning calculations and is assumed fixed. Floor space constraints, high stepper costs, and long delivery lead times for extra steppers ensure the validity of this assumption for the near-term. All other wafer fabrication operations are assumed to have excess capacity so that stepper capacity is the process constraint.

The FAB is modeled as a series of specific operations separated by transfer times. The operations of concern are the lithography patterning steps required for each layer. The transfer times account for all other wafer processing that takes place in between lithography patterning steps (doping, heat treating, testing, etc). Reducing the majority of FAB operations to transfer times is possible since all other activities outside of lithography only increase the overall time a wafer spends in the system, rather than actually constraining capacity.

Assuming an initial model of only one product, all of the wafers follow the same fixed sequence of operations. Wafer layers are classified as front-end (FE) layers and back-end (BE) layers. FE layers create the transistor features on the wafer substrate, while BE layers form the metal circuitry that connects the transistors. The wafer process flow is to go through the FE layers first, and then the BE.

Inputs

Lot Size (L - number of wafers): Wafers fabricated in groups of 25, which make up a lot. The number of wafers per lot is assumed constant for all lots.

Batch Size (B – number of lots): Lots are often "cascaded" or processed sequentially in batches to minimize idle and setup time between lots. All of the lots within a batch must require the same operation.

Line Yield (ρ): Wafer losses are assumed to be uniformly distributed over all operations, requiring fewer wafers to be processed at each subsequent step. Line yield is modeled due to its direct impact on processing capacity and simple integration. Die yield is a function of processing and is not included in this model.

Number of Layers (M): Specifically, the number of front-end (M_{FE}) and back-end (M_{BE}) layers that are patterned on the wafer surface.

Number of steppers (N)

Run Rates (RR – wafers/hour): A run rate is required for each layer type, which designates the amount of wafers that can be processed in an hour. Run rates represent raw processing time only (not including setup or calibration)

Qualification Time (QT – hrs/week): Each week this specified amount of time is required to qualify a stepper to be able to process a specific layer. This value represents the fixed costs of stepper flexibility by reducing processing capacity.

Maintenance/Setup Time (MT – hrs/event): After a stepper processes a batch, maintenance and calibration activities are required. This value also represents the setup time required as a stepper switches from processing one layer to another.

Failure Time (ω - %): The average percentage of time each week that a stepper is unavailable due to random failures. 1- is equivalent to the average percentage of time each week that a stepper is available for production activities.

Minimum Idle Time (δ - %): The minimum percentage of time each week that a stepper must be idle. This control variable limits the capacity of each stepper and is useful for modeling various levels of stepper utilization.

Constraints

Balanced Weekly Production: Sufficient stepper capacity must be allocated to each layer to balance weekly production.

Copper Segregation: Some of the back-end processes use copper, which is extremely mobile and easily damages the electrical properties of the semiconductor transistors that are fabricated during front-end operations.

Therefore, to prevent copper contamination in the FE operations, if a stepper has been assigned to a front-end (FE) layer, it can only process other FE layers.

Flexibility Requirements: In order to test various flexibility arrangements, minimum (or maximum) tool group flexibility requirements can be made.

Minimum number of Steppers per Layer (SL_{min}) provides a lower bound on the number of steppers that are required to be qualified to process a layer.

Minimum/Maximum number of Layers per Stepper (LS_{min} and LS_{max}) provides a lower or upper bound on the number of layers that a specific stepper must be qualified to process.

Average number of Steppers per Layer (RSL_{avg}) provides a strict constraint on the average SL for the entire tool group, as an alternative to individual min or max flexibility requirements.

The main model assumptions are:

- The total number of steppers is fixed.
- All other wafer fabrication operations have excess capacity.
- Only one product flow is considered.
- All steppers have equal processing capabilities (homogeneous tool group).
- Line Yield (but not die yield) is considered, and wafer losses are linearly distributed over all of the layers.

LP Formulation

A decomposition method was developed which makes general preliminary stepper allocations to the FE or BE operations [18]. This allows the LP to be separated into FE and BE problems, and with an iterative process the optimal number of FE and BE steppers are determined that maximizes total tool group throughput.

Decision Variables

Three types of decision variables are used:

WS = Wafer starts per week; the amount of wafers that begin processing each week.

X_{ij} = the fraction of time that stepper S_j dedicates to producing layer L_i (does not include setup, maintenance, or qualification),

$i = 1, \ldots, M$, $j = 1, \ldots, N$.

Z_{ij} = a binary variable which indicates the assignment of stepper S_j capacity to layer L_i, $i = 1, \ldots, M$, $j = 1, \ldots, N$. $Z_{ij} = \{1 \text{ if } X_{ij} > 0, 0 \text{ otherwise}\}$

For example:

	S_1	$\cdots$	S_j	$\cdots$	S_N
L_1	X_{11}	X_{12}	$\cdots$		X_{1N}
	X_{21}	$\ddots$			$\vdots$
L_i	$\vdots$		X_{ij}	$\cdots$	X_{iN}
			$\vdots$		$\vdots$
L_M	X_{M1}	$\cdots$	X_{Mj}	$\cdots$	X_{MN}

Where $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$

Calculated Values

APL_i = actual production for layer L_i (wafers/day) = $168(RR_{Li})\sum_{j=1}^{N} X_{ij}$

LS_j = the number of layers stepper S_j is qualified to process = $\sum_{i=1}^{M} Z_{ij}$

YPL = Yield per layer (%) = $10^{\frac{\log \rho}{M-1}}$

RPL_i = required production for layer L_i (wafers/day) = $WS(YPL)^{i-1}$

SL_i = the number of steppers qualified to process layer L_i = $\sum_{j=1}^{N} Z_{ij}$

SL_{avg} = the average number of stepper qualifications per layer = $\frac{\sum_{i=1}^{M}\sum_{j=1}^{N} Z_{ij}}{M}$

U_{Sj} = fraction of time each week that stepper S_j is utilized, including setups, maintenance, and qualification requirements.

$$= \sum_{i=1}^{M}\left(X_{ij}\left(1+\frac{RR_{Li}(MT)}{B(L)}\right)+\frac{Z_{ij}(QT)}{168}\right)$$

U_{max} = the maximum fraction of time each week that a stepper can be utilized.
= $1 - \omega - \delta$

Objective Function

The goal of the LP is to allocate steppers to layers in an arrangement that maximizes the throughput of the tool group. Utilization and flexibility constraints can then be adjusted to provide optimal solutions for a range of system scenarios. In this model, throughput (or output) is equal to WS * (ρ). Since is a constant:

Objective Function = Maximize WS

Constraints

The calculated values are used to form the following constraints:

Allocated stepper capacity must be greater than or equal to the required production for each week (168 hours).

$$APL_i \geq RPL_i \quad \rightarrow 168(RR_{Li})\sum_{j=1}^{N} X_{ij} \geq WS(YPL)^{i-1}$$

A stepper has limited capacity and cannot be utilized more than the max utilization requirement:

$$U_{Sj} \leq U_{max} \quad \rightarrow \sum_{i=1}^{M}\left(X_{ij}\left(1+\frac{RR_{Li}(MT)}{B(L)}\right)+\frac{Z_{ij}(QT)}{168}\right) \leq 1-\omega-\delta$$

If flexibility constraints are set, the following constraints ensure they are not exceeded:

$$LS_j \geq LS_{min\,j} \quad \rightarrow \sum_{i=1}^{M} Z_{ij} \geq LS_{min\,j}$$

$$LS_j \leq LS_{max\,j} \quad \rightarrow \sum_{i=1}^{M} Z_{ij} \leq LS_{max\,j}$$

$$SL_{avg} = RSL_{avg} \quad \rightarrow \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} Z_{ij}}{M} = RSL_{avg}$$

$$SL_i \geq SL_{min\,i} \quad \rightarrow \sum_{j=1}^{N} Z_{ij} \geq SL_{min\,i}$$

Furthermore,
$Z_{ij} \geq X_{ij}$
Z_{ij} is binary
$WS, X_{ij}, Z_{ij} \geq 0$ for all $i = 1,\ldots, M$ and $j = 1,\ldots, N$

Model Size and Solver Performance

The decomposed LP proved easy to solve over a wide range of inputs. By dealing only with the FE or BE allocations, the number of integer constraints was kept to a minimum. Table 1 shows the problem size, which is dependent on the number of layers (M) and number of steppers (N).

Table 1: Decomposed LP Problem Size

Decision Variables		Constraints	
Binary	MN	APLi $\geq$ RPLi	M
Continuous	MN+1	USj $\leq$ Umax	N
		LSj $\geq$ LS min j	N
		LSj $\leq$ LS max j	N
		$SL_{avg} = RSL_{avg}$	1
		SLi $\geq$ SL min i	M
		Zij $\geq$ Xij	MN
Total Variables = 2MN + 1		Total Constraints = MN + 3N + 2M + 1	

For example, a high-volume FAB with a FE tool group of 20 steppers and 15 layers corresponds to the following LP size:

M = 15, N = 20
Total Variables = 601 (Binary = 300; Continuous = 301)
Total Constraints = 391

It is also important to note that due to the inherent nature of this type of allocation problem, the LP generates multiple optimal solutions for any set of parameters. Some of the solutions are essentially identical, but each solution must be interpreted to ensure that the appropriate constraints are being considered.

Simulation Development

The LP model allocates steppers using static, average inputs and does not take into account the dynamics of the everyday randomness inherent in the lithography tool group. Simulation is necessary in order to begin to understand the real influence that flexibility has on FAB performance. By simulating the tool group using the same basic parameters and approach as with the LP, an equivalent model was generated that allows inputs to be varied under random distributions. The simulation can then test the LP solutions generated for each set of flexibility constraints, relating flexibility to the key performance metrics of throughput and cycle time.

The main stochastic parameters to be modeled by the simulation are:

- Exponential Mean Time Between Failures (MTBF) and a triangularly distributed average Time Down (TD) equate to the average % of failure time (ω) used in the LP, according to the equation:

$$1 - \omega = \frac{MTBF}{MTBF + TD}$$

Exponential mean time between failures is a common distribution for modeling machine failures [20].

- A triangular distribution (min, mean, max) was used to model transfer times between patterning operations. A triangular distribution for failure times was chosen because of its simplicity in matching general expectations of tool performance.

Without variation in transfer times, the discrete WIP re-entrant flow would feed directly from one layer to the next. These two parameters (stepper availability and wafer arrivals) are the fundamental sources of variability within the stepper tool group. In addition to these two random variables, several other inputs are required [21]. Table 2 lists these inputs and provides a definition for their use in the simulation. The simulation model used for this study mimics the LP test scenario of 4 steppers and 4 layers. However, the simulation can easily be scaled to full tool group size as time and computer capabilities permit [17, 20].

Table 2: Definitions of Simulation Model Inputs

Input	Definition
Number of Replications	The number of times that a simulation scenario is repeated.
Length of Replication (hrs)	The length of time that data is collected during each replication.
Warm up Period (hrs)	The length of time a replication is run before data is collected.
Time Between Arrivals (TBArriv - hrs)	The length of time between each new batch release into the FAB. TBArriv = L * B * 168 / WS
Batch Size (lots)	The number of lots in each batch.
Batch Process Time(hrs)	This value is reduced as the number of wafers in each lot decreases due to YPL.
Switch Setup Time (hrs)	Required when a stepper switches processing from one layer to another.
Qual Time/week (QT – hrs)	The amount of qualification time a stepper must spend each week for each layer it is qualified to process.
P (number batches)	The scheduling algorithm factor for the required difference in queue size (number batches) before layer priorities are switched.
Batch Setup Time (MT – hrs)	The length of time required for maintenance/calibration before a batch can be processed.
Downtime Time (hrs)	The length of time a stepper is down when a failure occurs.
Time Between Downtimes (hrs)	The length of time between each failure.
Transfer Time (hrs)	The length of time it takes a batch to travel from one layer to the next.

A basic scheduling policy was used to balance WIP between each layer. The algorithm for this "Largest Queue Rule" is as follows:

- If a wafer lot arrives at a queue and one stepper is available (and qualified to process it), then the stepper will process the wafer lot.
- If a wafer lot arrives at a queue and multiple steppers are available and qualified, then the lot will be processed on the stepper that is highest in its pre-determined preferred order. This order is set to maximize stepper-layer dedication.
- If a lot arrives at a queue, all steppers are busy, and multiple lots are waiting to be processed, then when a stepper becomes available the lot in the highest priority queue is processed first. In a tie, the lot with the longest queue time has priority.

Layer priorities are determined when a stepper processes a wafer lot. If the largest queue (Q_{max}) is P larger than the one the wafer just left (Q_{last}), then Q_{max} is assigned a priority of 1. Q_{last} is then assigned a lower priority of 2 if:

$$Q_{last} >= Q_{max} - 1.5*P$$

Otherwise, it is assigned the lowest priority of 3.

The effectiveness of the "Longest Queue Time" rule depends on the value of P and the dynamics of the system. Research has shown that when the steppers consistently process maximum batch sizes then relaxing a setup-avoidance scheduling policy can actually improve FAB performance [10]. This should be further evaluated in subsequent research.

The key to obtaining useful results from linear programming and simulation is to use both tools simultaneously. The following 9-step methodology decomposes the stepper tool group into FE and BE steppers, solves each LP to form a total solution, and then simulates the solution to obtain performance metrics:

Steps 1 to 4 – Preliminary Calculations: to determine the number of FE and BE steppers to be used for the decomposed LP.

1. The following inputs are required:

M_{FE}, M_{BE} ($M_{FE}+M_{BE} = M$)	δ
N	MT
ρ, ρ_{FE}	QT
RR_{FE}, RR_{BE}	B
ω	L

2. Calculate WS_{max} and set $WS_{prelim} = WS_{max}$:

Let RRmax = maximum $\{RR_{FE}, RR_{BE}\}$
RRmin = minimum $\{RR_{FE}, RR_{BE}\}$

$$WS_{max} = \frac{168N(RR_{max})(1-\omega-\delta)-QT(N)(RR_{min})}{M(\rho)\left[\frac{MT(RR_{min})}{B(L)}+1\right]}$$

3. Use WS_{prelim} to calculate N_{FE} and N_{BE} and round up to the next integer:

$$NFE = \frac{\sum_{i=1}^{M_{FE}} RPL_i}{168(1-\omega-\delta)-QT}\left(\frac{1}{RR_{FE}}+\frac{MT}{B(L)}\right)$$

$$NBE = \frac{\sum_{i=M_{FE}+1}^{M} RPL_i}{168(1-\omega-\delta)-QT}\left(\frac{1}{RR_{BE}}+\frac{MT}{B(L)}\right)$$

4. Check WS_{prelim} feasibility:

If $N_{FE} + N_{BE} > N$, then set $WS_{prelim} = WS_{prelim} * (0.99)$, and repeat at step 3. 0.99 is a subjective decrementing value and can be increased or decreased to refine the change in WS_{prelim} between iterations.

If $N_{FE} + N_{BE} \leq N$ then continue to 5.

Steps 5 to 7 – LP Solution: to solve FE, BE stepper allocations and obtain optimal total solution. N_{FE} is adjusted (incremented or decremented) to balance capacity between FE and BE tool groups.

5. Solve FE LP using N_{FE} to find WS.

6. Solve BE LP using $N_{BE} = N - N_{FE}$ to find WS_{BE}.

7. If $WS(\rho_{FE}) > WS_{BE}$ then set $N_{FE} = N_{FE} - 1$ and repeat at step 5.
   If $WS(\rho_{FE}) < WS_{BE}$ then set $N_{FE} = N_{FE} + 1$ and repeat at step 5.
   If $WS(\rho_{FE}) \approx WS_{BE}$ then continue to 8.

Steps 8 and 9 – Simulation: to stochastically model the LP solutions and obtain FAB performance metrics.

8. Using LP solution of stepper-layer assignments, run the simulation to obtain performance metrics:

   Throughput = Batches Out/Batches Started for simulated period
   Cycle Time = Avg CT/Base CT
   Throughput Variability = CT and Outs Range

9. Set new SL and LS flexibility constraints and repeat at 5. Continue iteration over desired range of flexibility constraints to obtain performance curve.

Results and Discussion

A baseline scenario was used for testing and validating the LP and Simulation methodology. The data used in this baseline case is listed in Table 3 and Table 4. These input values are selected based on commonly observed values experienced in a particular FAB lithography tool group.

Table 3: Baseline LP Input Values

LP Model Inputs	Baseline Value	LP Model Inputs	Baseline Value
A_{min}	0 Wafers	ρ	90%
ω	20%	M	4 Layers
δ	5%	N	4 Steppers
B	4 Lots	MT	0.5 hours
L	25 Wafers	QT	1.5 hours
LS and SL min, max	Variable	RR	30 Wafers/hr

Table 4: Baseline Simulation Input Values

Simulation Inputs	Baseline Value
Number of Replications	15
Length of Replication (hours)	720
Warm up Period (hours)	3000
Time Between Arrivals (TBArriv)	Variable
Batch Process Time for Layer 1 (hrs)	3.33
Batch Process Time for Layer 2 (hrs)	3.22
Batch Process Time for Layer 3 (hrs)	3.11
Batch Process Time for Layer 4 (hrs)	3.00
Switch Setup Time (hrs)	0.5
Stepper and Layer Flexibility (SL & LS)	LP Input
P (number batches)	3
Downtime Time (TS - hrs)	Triangular (4,12,20)
Time Between Downtimes (MTBF - hrs)	Exponential (48)
Transfer Time (hrs)	Triangular (5,10,15)

Using the values in Table 4, the LP generates stepper-layer qualifications for average SL = 1, 1.5, 2, 3, and 4 levels of flexibility, with SL = LS (all of the steppers and layers are equally flexible). For the SL = 1.5 arrangement, the LP assigns SL = 2 to layers 1 and 2, and SL = 1 to layers 3 and 4 to make the average SL = 1.5. The line yield assumption causes layer 1 to require the most capacity and layer 4 require the least, and so the LP balances the line by assigning more capacity to lower numbered layers.

Flexibility and FAB Performance

An important measure of cycle time is the cycle time ratio (CTR, also known as the multiplier of theoretical cycle time) [7, 11]. CTR is the ratio of the average cycle time of a batch to the base cycle time. Base cycle time is defined as the average cycle time with no waiting. Figure 1 shows the CTR as each flexibility level is simulated under a range of WS loadings. The LP maximum WS results for 0% and 15% idle times are also shown in Figure 9 as the vertical dotted lines.

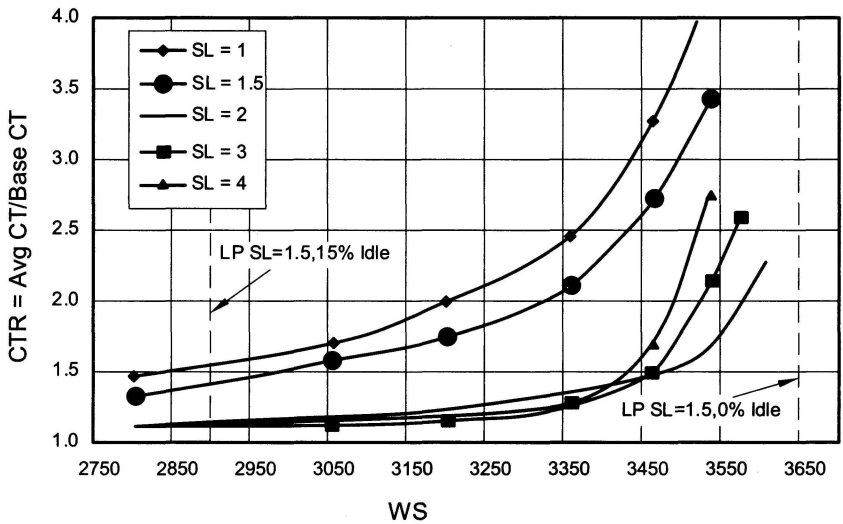

Figure 1: CTR per WS and Flexibility Level. These Simulation Results Show the CTR for Each Level of SL Flexibility Over a Range of WS Values. The Vertical Dotted Lines Show the Corresponding LP Maximum WS Results for 0% and 15% Stepper Idle Times.

Figure 1 clearly shows a significant decrease in CTR as flexibility is introduced into the tool group at any WS level. This is consistent with results of [8] and [17]. However, an important additional insight is that flexibility does

not always increase the CTR at each WS level. At lower WS levels (and correspondingly lower stepper utilization), the biggest gain in CTR reduction is achieved by increasing flexibility to SL = 2. Additional flexibility does not yield significant improvements in the CTR. In fact, at higher WS levels (higher tool utilization) stepper tool groups with SL = 3 or 4 perform worse than the tool group with SL = 2. This decrease in performance is because the extra flexibility at these levels does not justify the extra setup and qualification capacity costs. At lower WS levels the added capacity costs of flexibility (QT and ST) are absorbed by idle time. As WS is increased, idle time decreases, until eventually the capacity costs of added flexibility use all of the available idle time.

Another important result shown in Figure 1 is that at lower WS (lower tool utilization) the benefits from flexibility are less significant than at higher WS. For example, at WS = 2900, average idle time = 15% and increasing SL from 1 to 2 decreases CTR from 1.55 to 1.2 (a change of 0.35). But at WS = 3450, average idle time = 4% and the same increase in SL decreases CTR from 3.5 to 1.5 (a change of 2.0)! This demonstrates the significant CTR benefits that an appropriate level of flexibility (in this case, SL = 2) can have when tools are utilized at high levels.

The dotted vertical LP optimal lines in Figure 1 provide additional insight. For comparison purposes, the LP was solved using a range of idle time (δ) minimum requirements. The δ = 0% and 15% lines are shown, and both were optimal for a flexibility of SL = 1.5. The δ = 0% line is interpreted as the maximum feasible WS that the stepper tool group can maintain given the specified parameters. The SL = 2 level of flexibility comes close to this maximum WS at a low CTR. This shows that SL = 2 flexibility easily compensates the random availability of the steppers and the randomness in arrivals (transfer times). Although this system is not as complex or variable as full-sized tool group, these results reveal an important potential benefit of increasing flexibility from a SL =1 (totally dedicated) scenario or decreasing flexibility from a SL = 4 (totally flexible) scenario.

Flexibility and Throughput

Flexibility increases tool group capacity by allowing machine capacity to be more evenly distributed among all of the tasks the tool group must perform. Cycle-time (CT) constrained throughput is the maximum amount of throughput that a system can provide given a specified average cycle time, and is an excellent metric for measuring the throughput impact of different flexibility arrangements [7]. From the simulation results outlined above, Figure 2 shows the CT-constrained throughput in terms of WS for cycle times of 1.5, 2, and 3. This method of representing the results shows throughput gains by increasing flexibility to the optimal level of SL = 2, and then deteriorating capacity as it is increased beyond that point (due to increased setup and qualification requirements).

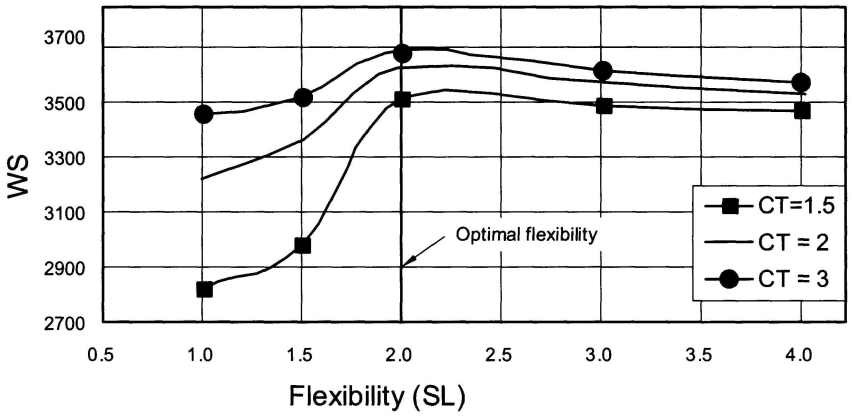

Figure 2: CT-Constrained Throughput. WS Values are Shown at Different Levels of SL Flexibility and CT Constraints. The Optimal Flexibility Arrangement of SL = 2 is Highlighted.

Flexibility and Throughput/CT Variability

The variability of fabricated wafer output can be another critical measure of FAB performance since production is often geared towards weekly customer commitments. Results show that increased flexibility can correspond to less variable output, but at high levels of utilization, too much flexibility can actually increase variability. As a first cut estimation of variability, the range of the average monthly CTR was calculated using 15 independent simulations. Figure 3 shows the correlation that increasing stepper group flexibility (at least to SL = 2) significantly decreases variability, while increasing flexibility to SL = 3 and 4 increases variability at higher WS levels. For clarity, SL = 4 is not shown on the graph. A more thorough analysis of throughput variability is warranted for a complete analysis of tool group flexibility. For example, cycle time standard deviation of weekly throughput statistics is an excellent measure of variability [11].

A consistent result demonstrated by both Figure 1 and Figure 3 is that the benefits of the right level of tool group flexibility are largest at high tool utilization (high WS values). By the performance metrics of CTR and CTR Range, SL = 2 is clearly the best arrangement when tool utilization is maximized. As idle time increases, the difference in performance between the flexibility arrangements is less clear. However, the stepper tool group usually experiences high utilization because it is the process constraint. Therefore, the optimal level of flexibility can be clearly identified using this combined LP/simulation methodology.

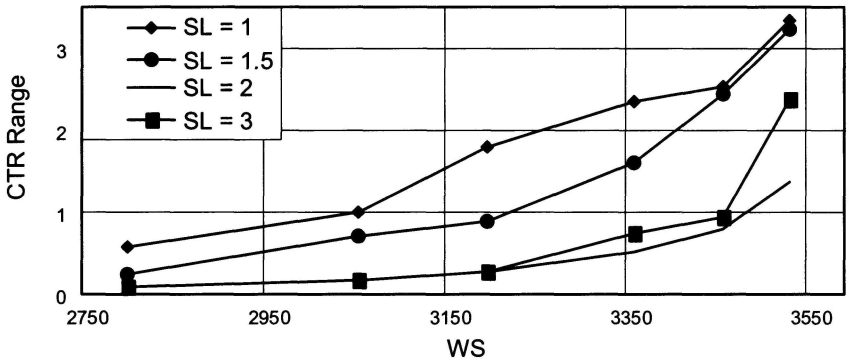

Figure 3: Range of Average CTR. The Range of CTR (Max CTR – Min CTR) for Each Set of Simulation Experiments is Plotted per WS and Flexibility Level. SL = 4 is not Shown for Clarity.

Failure/Downtime Frequency (MTBF) and Downtime Duration (TD)

Figure 4 shows the simulation results as MTBF is increased to 96 hours and decreased to 24 hours from the baseline scenario of 48 hours. TD is also adjusted according to the relationship:

$1-\omega = \frac{\text{MTBF}}{\text{MTBF}+\text{TD}}$ so that all cases have the same long-term availability of 80% (= 20%).

When compared to Figure 1, the sensitivity analysis results in Figure 4 show that as MTBF is increased (indicated by the bold lines) the SL = 1 tool group arrangement performs much worse. The SL = 2 and 3 arrangements are much more robust, but still increase in CTR from about 1.3 to 1.5 at WS = 3350. Alternatively, as MTBF is decreased, the SL = 1 tool group arrangement performs much closer to the other flexibility arrangements than in the baseline scenario. This illustrates that the less flexible arrangement is much more sensitive to changes in MTBF availability approximations, and is much more susceptible to uncertain stepper availability.

At a constant long-run average failure time percentage (ω), if MTBF decreases then TD becomes shorter (more frequent, shorter failures). When failures are consistent, frequent, and shorter, the need for flexibility to maintain a given CTR is reduced. On the other hand, as MTBF increases then TD becomes longer for a given value of ω, and failures become more variable, less frequent, and longer. Without a change in flexibility, the CT for a given level of throughput increases dramatically with increased MTBF, as queues become much larger while waiting for tools to become available.

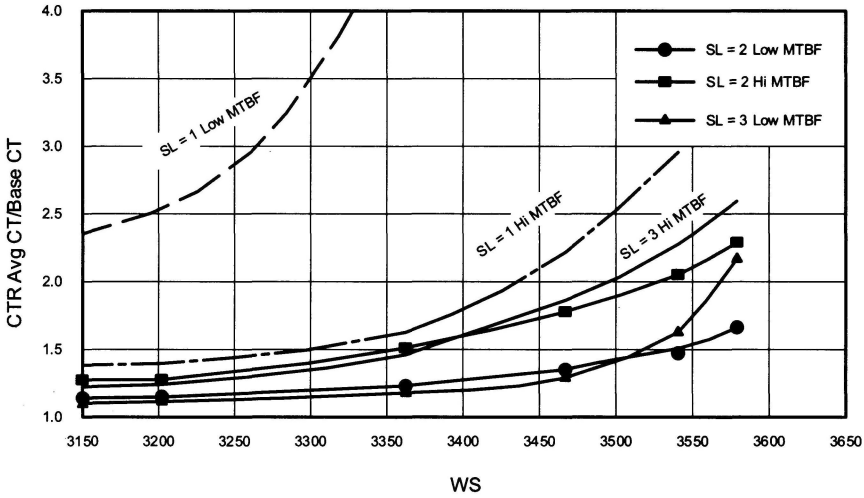

Figure 4: CTR and Flexibility for High/Low MTBF. This Sensitivity Analysis Shows the CTR When MTBF is Changed to a High Value of 96 Hours and a Low Value of 24 Hours from the Baseline of 48 Hours. = 20% for All Cases.

Qualification and Setup Time

Qualification and setup times are the model parameters that define the capacity costs of flexibility. Qualification time (QT) essentially associates a fixed stepper capacity cost with increased flexibility. As QT becomes large, the tool groups with high levels of flexibility have significant reductions in available processing capacity, while the SL = 1 arrangement is minimally affected. A simulation sensitivity analysis was conducted by changing QT is changed to a high value of 3 hours per week and a low value of 0.5 hours per week from the baseline value of 1.5 hours.

The results of the analysis show while SL = 1 performance changes minimally with QT changes, the SL = 3 CTR performance changes dramatically. In the low QT scenario, the CTR for SL = 3 stays low (< 1.5) until WS exceeds 3540. In the high QT scenario, the CTR for SL = 3 quickly exceeds 1.5 as WS increase past 3350. Alternatively, SL = 2 proves to have superior CTR performance under the high QT scenario and at high WS values. In the low QT scenario, it is essentially equivalent to the higher flexibility arrangement. Again, this demonstrates the importance of using the LP and simulation methodology in order to find an optimal flexibility level that is robust to changes in system parameters while still yielding significant improvements in FAB performance.

Setup time (ST) can also significantly alter the amount of time it takes to process a wafer. As setup time increases it becomes more costly to switch from one layer to another, limiting the benefits of flexibility. A simulation sensitivity analysis was conducted by changing ST to a high value of 2 hours and a low value of 0.1 hours from the baseline value of 0.5 hours. The results are very similar to the QT sensitivity analysis. In essence, the more flexible tool group (SL = 3) loses significant processing capacity at high ST values and CTR quickly increases with higher WS values. The CTR for the SL = 2 arrangement does increase at the higher ST, but remains the optimal choice in all of the scenarios based on CTR performance. It should be noted that the no-flexibility arrangement (SL = 1) is unaffected by ST changes.

Conclusions

This paper develops an effective methodology of using LP and simulation in an integrated manner to determine the relationship that lithography tool group flexibility has on FAB performance in semiconductor manufacturing.

- A deterministic linear model was developed that allocates steppers to layers given a set of defining parameters and constraints. The LP makes appropriate stepper-layer qualifications to maximize throughput.
- A simulation was developed that tests each LP solution under stochastic conditions to determine FAB performance according to cycle time and throughput metrics.
- A methodology was proposed that integrates the use of the LP and simulation. The methodology uses an iterative analysis technique to define the relationship between the performance and tool group flexibility.
- Initial results were obtained as the methodology was verified with a simplified model of the lithography tool group. The results show that a flexibility level of SL = 2 optimized the CTR and CT - Constrained throughput.
- Sensitivity analysis was carried out to identify the important simulation model parameters. Changes in MTBF, QT, and ST had significant effects on each tool group configuration.

Acknowledgement

The authors would like to acknowledge the critical comments and discussion received from Chris Hans, Intel Corporation, Portland, and T. Klostorin, Department of Management Science, School of Business, University of Washington.

References

[1] Van Zant, P. (2000). Microchip Fabrication, Fourth Edition, McGraw-Hill, Boston.

[2] Kumar, P.R. (1993). "Re-Entrant Lines", Queuing Systems: Theory and Applications: Special Issue on Queuing Networks, Vol. 13, 87-100.

[3] Robinson, Jennifer, K. (1998). "Capacity Planning in a Semiconductor Wafer Fabrication Facility with Time Constraints Between Process Steps," Doctor of Philosophy Dissertation, Department of Engineering, University of Massachusetts Amherst.

[4] Uzsoy, R., Lee, C. and Martin-Vega, L. (1992). "A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part I: System Characteristics, Performance Evaluation, and Production Planning," IIE Transactions, Vol. 24, No. 4, 47-60.

[5] Flack, W., Flores, G., Walther, A. and Ferreira, M. (September 1994). "A Reticle Correction Technique to Minimize Lens Distortion Effects," 14th Annual BACUS Symposium on Photomask Technology, Santa Clara, CA.

[6] Gupta, D. (1993). "On Measurement and Valuation of Manufacturing Flexibility", International Journal of Production Research, Vol. 31, No. 12, 2947-2958.

[7] Fowler, J., Brown, S., Gold, H. and Schoemig, A. (October 1997). "Measurable Improvements in Cycle-Time Constrained Capacity," Proceedings of the 6th International Symposium on Semiconductor Manufacturing, San Francisco, A21-A24.

[8] Hase, R., Takoudis, C. and Uzsoy, R. (1997). "Alternative Facility Layouts for Semiconductor Wafer Fabrication Facilities," IEEE Transactions on Components, Packaging, and Manufacturing Technology-Part C, Vol. 20, No. 2, 152-163.

[10] Peikert, A., Thoma, J. and Brown, S. (1998). "A Rapid Modeling Technique for Measurable Improvements in Factory Performance," Proceedings of the 1998 Winter Simulation Conference, 1011-1015.

[11] Lu, S., Ramaswamy, D. and Kumar, P. (1994). "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants," IEEE Transactions on Semiconductor Manufacturing, Vol. 7, No. 3, 374-388.

[12] Mittler, M. (1996). "Two-Moment Analysis of Alternative Tool Models with Random Breakdowns," Proceedings of the 1996 IEEE Conference on Emerging Technologies and Factory Automation, Kauai, HI, 546-552.

[13] Gupta, D., Gerchak, Y. and Buzacott, J. (1992). "The Optimal Mix of Flexible and Dedicated Manufacturing Capacities: Hedging Against Demand Uncertainties," International Journal of Production Economics, Vol. 28, No. 3, 309-319.

[14] Benjaafar, S. and Gupta, D. (1999). "Workload Allocation in Multi-Product, Multi-Facility Production Systems with Setup Times," IIE Transactions, Vol. 31, No. 3, 339-352.

[15] Leachman, R. and Carmon, T. (1992). "On Capacity Modeling for Production Planning with Alternative Machine Types," IIE Transactions, Vol. 24, No. 4, 62-72.

[16] Rohan, D. (1992). "Resource Sharing in Capacity Analysis," IEEE/SEMI Advanced Semiconductor Manufacturing Conference, 39-42.

[17] Rohan, D. (1999). "Machine Dedication Under Product and Process Diversity," Proceedings of the 1999 Winter Simulation Conference, 897-902.

[18] Ovacik, I. and Uzsoy, R. (1997). "Decomposition Methods for Complex Factory Scheduling Problems," Kluwer Academic Publishers, Boston.

[19] Lindo Systems Incorporated Internet Site: http://lindo.com/cgi/frameset.cgi?leftwb.html;wbf.html

[20] Hillier, F. and Lieberman, G. (2001). Introduction to Operations Research, McGraw-Hill, Boston.

[21] Kelton, D., Sadowski, R. and Sadowski, D. (1998). Simulation with Arena, McGraw-Hill, Boston.

JOURNAL OF MECHANICAL ENGINEERING (JMechE)

Aims & Scope

Journal of Mechanical Engineering (formerly known as Journal of Faculty of Mechanical Engineering), is an international journal which provides a forum for researchers and academicians worldwide to publish the research findings and the educational methods they are engaged in. This Journal acts as a vital link for the mechanical engineering community for rapid dissemination of their academic pursuits as well as a showcase of the research activity of FKM for the outside world.

Contributions are invited from various disciplines that are allied to mechanical engineering. The contributions should be based on original research works. An attempt will be made to review the submitted contributions with competent internal and external reviewers as to the suitability of the paper for satisfying the objectives of the journal.

All papers submitted to JMechE are subjected to a rigorous reviewing process through a worldwide network of specialized and competent referees. To be considered for publication, each paper should have at least two positive referee's assessments.

General Instructions

Manuscripts intended for publication in the Journal FKM should be written in camera ready form with production-quality figures and done electronically in Microsoft Word 2000 (or above) and submitted along with one hard copy. Manuscripts should be typed using Times New Roman font (point 10) on one side of the paper (A4 size), single-spaced, with wide margins (1 in left and right, and 1 in on top and bottom). The manuscript should include the title of the paper; the author's name and affiliation; a short abstract of between 200 and 300 words which clearly summarises the paper; and a list of keywords. Limit your submission to a maximum of 20 typed pages.

Keywords

Keywords supplied by the author should appear on a line following the abstract. The keywords selected should be comprehensive and subject specific. Maximum of five keywords should be sufficient to cover the major subjects of a given paper. General terms should not appear as keywords, as they have little use as information retrieval tools. Please choose keywords to be as specific as possible and list the most specific first, proceeding to the most general last.

Units

All scientific and technical data presented should be stated in SI units.

Footnotes

Footnotes should be kept to an absolute minimum and used only when essential.

Formulae

Formulae should be typewritten using MS Word compatible Equation Editor.

Tables

Tables, should be included within the text where appropriate and must be numbered consecutively with Arabic numerals and have titles that precede the table. They should be prepared in such a manner that no break is necessary.

Figures

Authors should appreciate the importance of good-quality illustrations. All graphs and diagrams should be referred to, for example, Figure 1 in the text. All figures must be numbered consecutively with Arabic numerals. A detailed caption should be provided below each figure according to the following format:

Figure 1: (a) A simple 2-D cantilever and (b) microcantilever with a diamond probe.

Figures should be embedded within the text where appropriate. Glossy photographs when required should be scanned to a resolution suitable with the reproduction requirements (1200 dpi generally will be sufficient).

References

Use squared brackets to indicate reference citation such as [1], [3]-[5] in the main text. Include references at the end of the paper according to the citations order that appears in the paper using the following format.

[1] M. K. Ghosh and A. Nagraj, "Turbulence flow in bearings," Proceedings of the Institution of Mechanical Engineers 218 (1), 61 - 64 (2004).

[2] H. Coelho and L. M. Pereira, "Automated reasoning in geometry theorem proving with Prolog," J. Automated Reasoning 2 (3), 329-390 (1986).

[3] P. N. Rao, Manufacturing Technology Foundry, Forming and Welding, 2nd ed. (McGraw Hill, Singapore, 2000), pp. 53 – 68.

[4] Hutchinson, F. David and M. Ahmed, U.S. Patent No. 6,912,127 (28 June 2005).